Домашнее задание № 5

Регрессионный анализ. Анализ выживаемости.

Крайний срок сдачи: 18 мая 2023 г., 23:59.

Каждое задание оценивается в 2 балла.

1. Рассмотрим простейшую линейную регрессию

$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1..n,$$

где ε_i - i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ и $x_1, ..., x_n$ - детерминированные точки, хотя бы две из которых различны. Обозначим через $\hat{\alpha}, \hat{\beta}$ оценки параметров α, β, полученные методом наименьших квадратов

$$\big(\hat{\alpha}, \hat{\beta}\big) := \arg\min_{\alpha,\beta} \mathcal{R}(\alpha, \beta), \quad \text{где} \quad \mathcal{R}(\alpha, \beta) = \sum_{i=1}^{n} \big(y_i - \alpha - \beta x_i\big)^2$$

Докажите, что эти оценки имеют нормальное распределение и являются несмещёнными оценками параметров α и β.

2. База данных PimaIndiansDiabetes2 содержит данные о различных медицинских показателях 768 жителей Индии. Информация о базе доступна по ссылке
`https://rdrr.io/cran/mlbench/man/PimaIndiansDiabetes.html`
Целью этого упражнения является анализ взаимосвязи переменной diabetes (индикатор болезни диабетом) с остальными переменными.

   (i) При помощи инструментов weights of evidence и information value разделите все переменные, кроме diabetes, на 3 группы (используя diabetes как event), и определите переменные, для которых information value превышает 0.4.

   (ii) Перекодируйте выбранные на предыдущем шаге переменные в соответствии с полученными группами. Исследуйте зависимость каждой из этих переменных с фактом наличия диабета при помощи анализа соответствующих таблиц сопряженности.

(iii) Для описания зависимости переменной diabetes от всех остальных переменных постройте модели логистической регрессии по всем переменным и отдельно по каждой перекодированной переменной. Определите, какая из построенных моделей лучше, используя ROC AUC.

3. База данных veteran содержит информацию о пациентах, больных раком лёгких, см.
`https://r-data.pmagunia.com/dataset/r-dataset-package-survival-veteran`
Данные находятся в пакете survival языка R.

   (a) Постройте кривую выживаемости Каплана-Мейера в зависимости от типа лечения (переменная trt[1]). При помощи логрангового теста определите, можно ли считать эти два метода одинаково эффективными.

   (b) Постройте регрессионную модель Кокса на основе всех переменных. Попробуйте улучшить качество модели при помощи следующих методов:

   - выделите выбросы по переменной time (время наблюдения) при помощи диаграммы размаха и постройте модель Кокса только на основе наблюдений, которые не являются выбросами;
   - выделите наиболее значимые переменные и постройте новую модель только на основе этих переменных.

4. Рассмотрим базу данных "LifeCycleSavings"(`https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/LifeCycleSavings.html`), содержащую информацию о среднем коэффициенте персональных сбережениях жителей 50 стран. Этот коэффициент для конкретного жителя вычисляется как отношение его совокупных личных сбережений к располагаемому доходу. Согласно гипотезе Модильяни, среднее по стране значение этого коэффициента зависит от

   - процента населения моложе 15 лет (LifeCycleSavings$pop15);

[1]Trt - это сокращение от treatment.

- процента населения старше 75 лет (LifeCycleSavings$pop75);
- располагаемого дохода на душу населения (LifeCycleSavings$dpi);
- процентной скорости изменения располагаемого дохода на душу населения (LifeCycleSavings$ddpi).

Представленные данные являются усреднёнными показателями за 1960–1970 гг.

(i) Для переменных "sr"(как y-переменной) и "pop15" (как x-переменной) постройте ядерную оценку регрессии при различных вариантах выбора ядра (гауссовское ядро и ядро Епанечникова) и различных методах выбора параметра bandwidth (критерий Акаике, обобщённый метод кросс-проверки). Найдите наилучший метод в смысле наименьшей среднеквадратичной ошибки.

(ii) Повторите вычисления для остальных трёх объясняющих переменных вместо "pop15". Выберите 2 переменные, которые по Вашему мнению наилучшим образом объясняют коэффициент персональных сбережений (в дальнейшем эти переменные будем называть V1 и V2). Объясните свой выбор.

(iii) На основе V1 и V2 постройте многомерную регрессию методом LOESS и линейную регрессию. Разделите случайным образом все страны на 2 группы: в одну группу отнесите примерно 80 % стран, в другую - 20 %. Оцените параметры модели LOESS и линейной регрессии по большей группе и проверьте качество моделей по меньшей. Выясните, какая из построенных моделей является более точной.

5. Пусть дан набор точек $(x_i, y_i), i = 1..n$. Для описания регрессионной зависимости между y_i и x_i будем использовать оценку Надарая-Ватсона

$$\hat{r}(x) = \frac{\sum_{i=1}^{n} y_i K\big((x - x_i)/h\big)}{\sum_{i=1}^{n} K\big((x - x_i)/h\big)}$$

с треугольным ядром

$$K(x) = (1 - |x|) \cdot \mathbb{I}\{|x| \leq 1\}$$

и параметром $h > 0$. Для случая $n = 6$ и $x_i = i,\ \forall i = 1..6$, вычислите сглаживающую матрицу L и эффективное количество степеней свободы (след матрицы L), если

(i) $h = 1/2$;

(ii) $h = 3/2$.

Комментарий. Напомним, что слаживающая матрица L - это такая матрица, что

$$\hat{\vec{y}} = L\vec{y},$$

где $\vec{y} = (y_1, ..., y_n)^\top, \hat{\vec{y}} = (\hat{r}(x_1), ..., \hat{r}(x_n))^\top$.